Weakly Supervised Object Detection

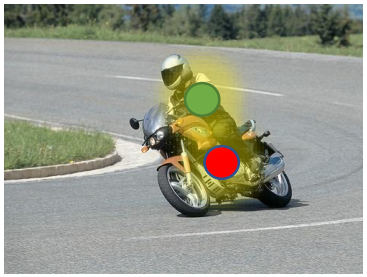
ECCV 2020
Weakly Supervised Learning in Computer Vision
Tutorial

Hakan Bilen
University of Edinburgh

Manual supervision for object recognition

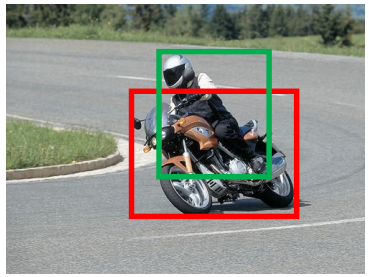
{motorbike (image-label), person (image-level)}

1 sec
per class



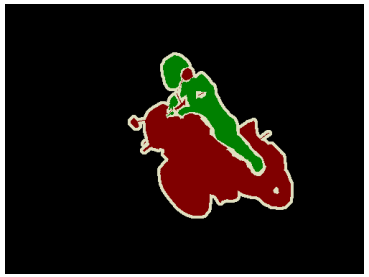
{motorbike (point), person (point)}

2.4 sec
per instance



{motorbike (b-box), person (b-box)}

10 sec
per instance



{motorbike (pixel labels), person (pixel labels)}

78 sec
per instance

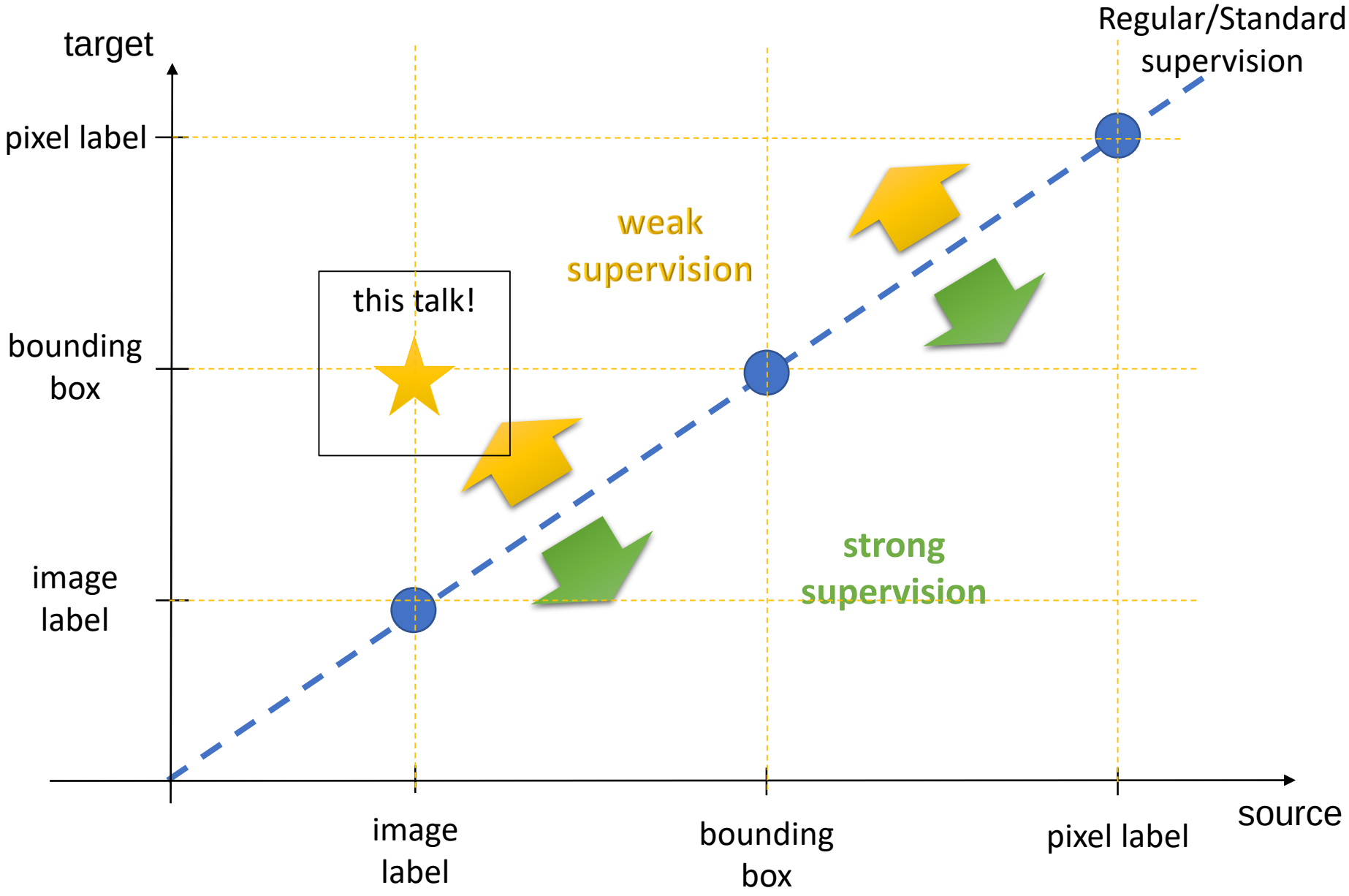
annotation time

Berman et al., What's the Point: Semantic Segmentation with Point Supervision, ECCV 16

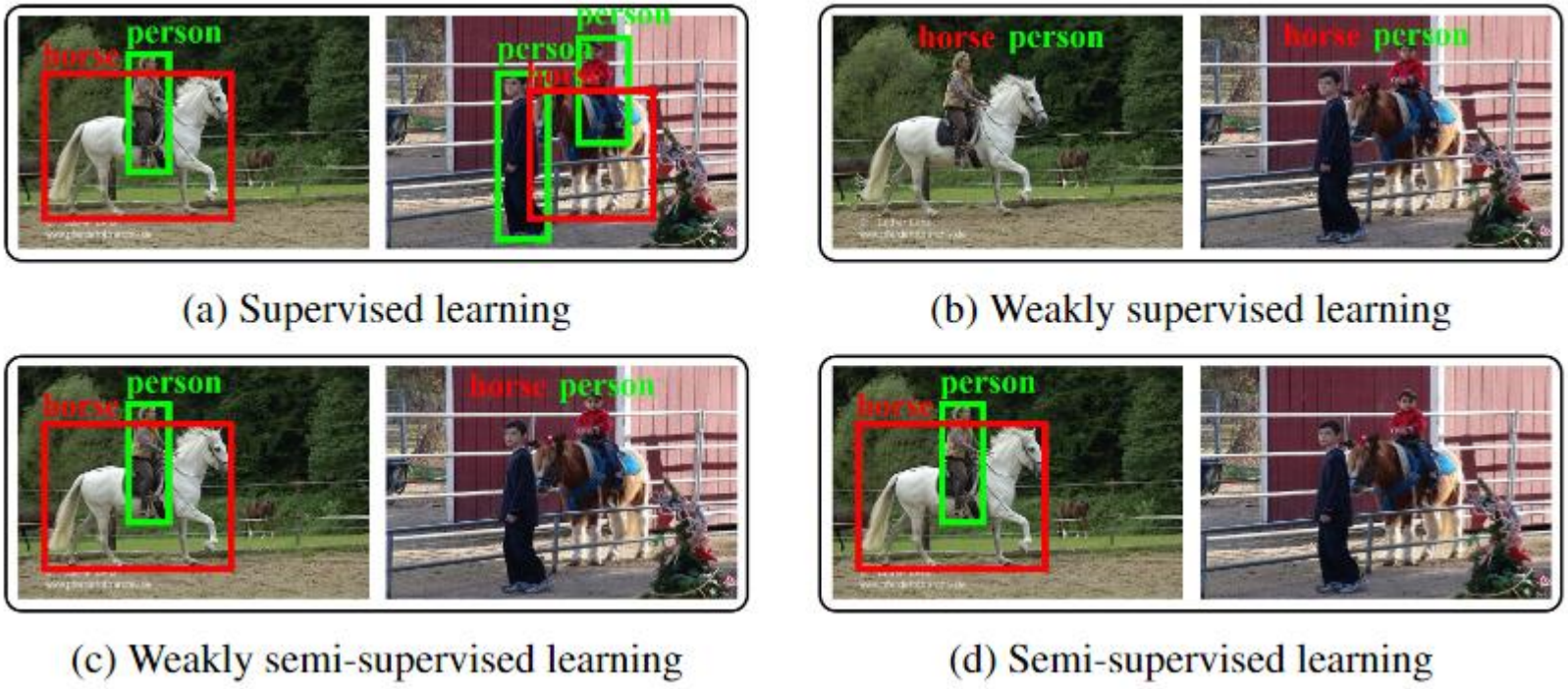
Weak supervision

Lower degree (or cheaper) annotation at train time than the required output at test time

Manual supervision for object recognition



Related topics: semi-supervised object detection



(a) Supervised learning

(b) Weakly supervised learning

(c) Weakly semi-supervised learning

(d) Semi-supervised learning

- Semi-supervised: learns from a set of labeled (bounding box) and unlabeled images

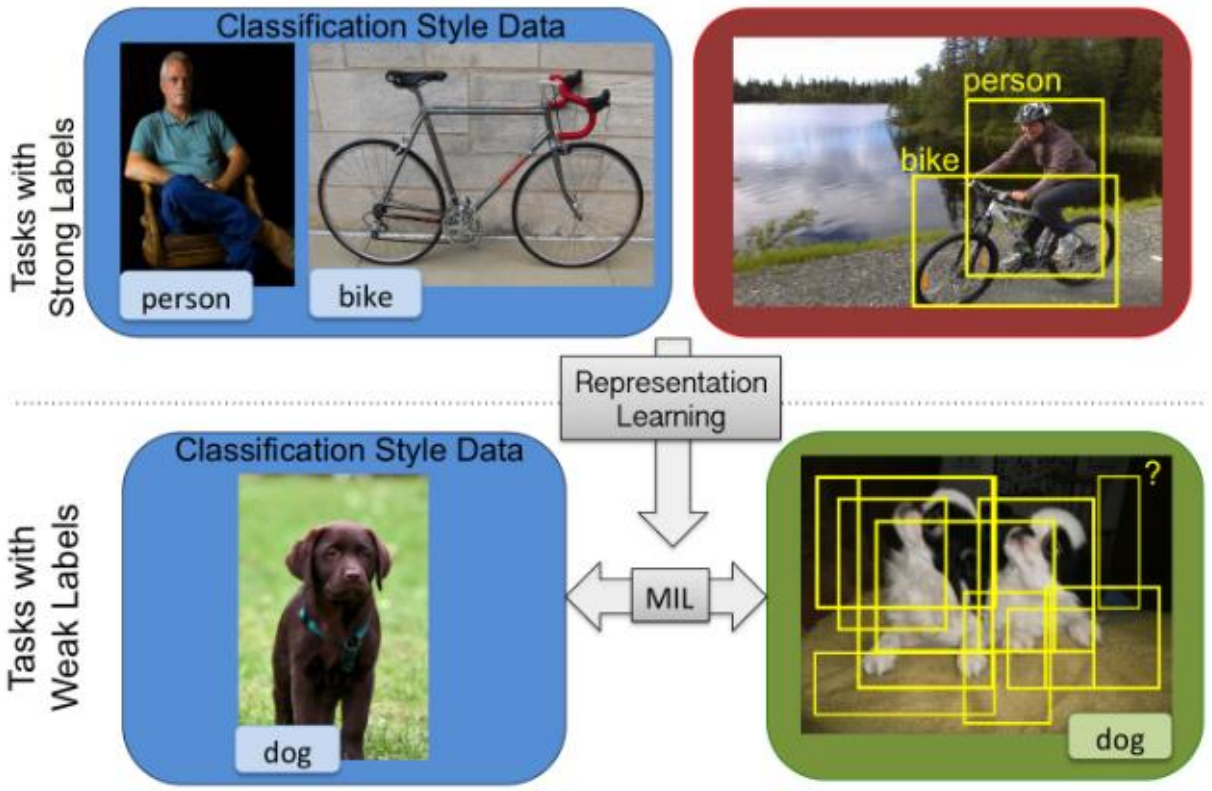
  [Wang CVPR 19]

- Weakly semi-supervised: learns from a set of labeled (bounding box) and weakly labeled (image label) images

  [Yan arXiv 1702.08740]

Image credit: [Jeong 19 NeurIPS]

Related topics: transfer learning



- Learns detectors for categories with only weak (image-level) labels
- Transfers the knowledge from auxiliary categories with bounding box labels

[Hoffman NIPS 14], [Tang CVPR 16]

Image credit: [Hoffman NIPS 14]

Related topics: Self-supervised learning

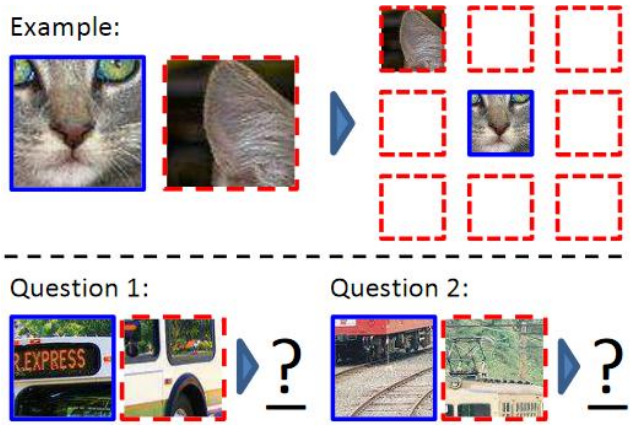


Image credits: [Doersch, ICCV 15]

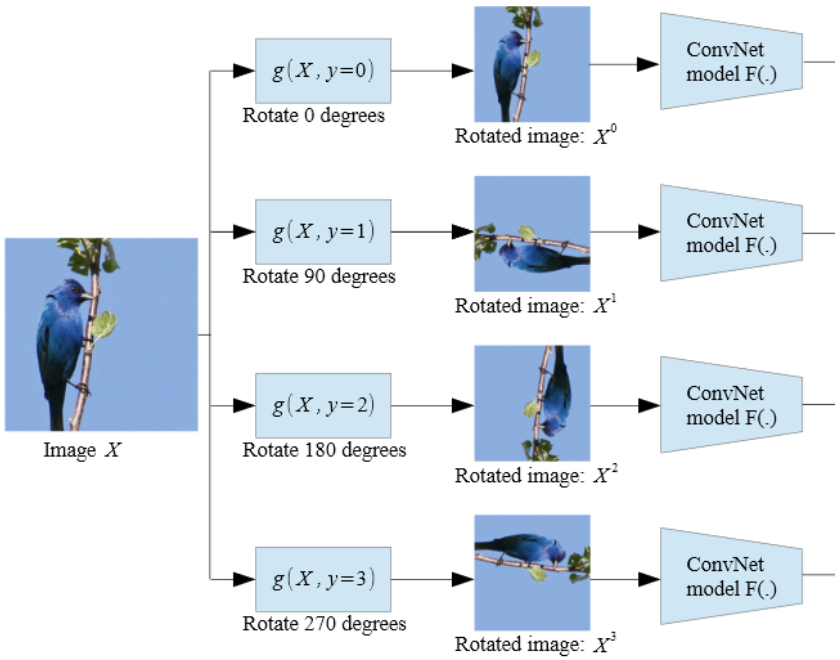


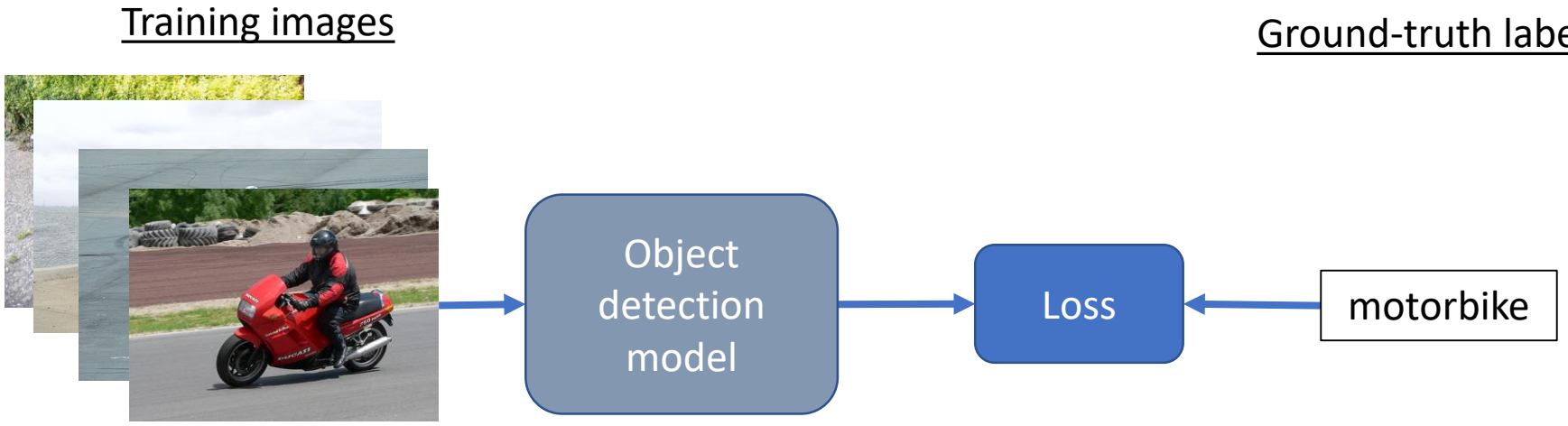
Image credits: [Gidaris ICLR 18]

- Defines a pretext task (eg predicting relative position of image patches, rotation of images) which does not require any manual annotation
- Pre-trained weights are transferred to other computer vision tasks such as object detection and semantic segmentation

- Requires regularly annotated data (eg bounding boxes for object detection) in the fine-tuning stage
- WSL typically uses pretrained networks
- Complementary to weakly supervised learning

Challenges

Weakly supervised object detection (WSOD)



What can we say at minimum?

1- When image is positive, at least one object instance from target category is present

2- When image is negative, no object instance from target category is present

Assumptions

1- [Classification] There exists a set of features present in positive images and absent in negative images

2- [Detection] The same features are only present in the (whole) target object instances

Generic Object Recognition

Intra-class variations

- Appearance
- Viewpoint
- Scale
- Aspect ratio

Background clutter

Occlusions



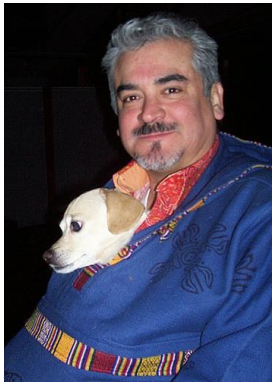



Weakly Supervised Learning

Parts vs object

"The same features are only present in the (whole) target object instances"



Question: What is a person?

a) Face

b) Face + upper body

c) Face + whole body ☺

Weakly Supervised Learning

Context vs object

"The same features are only present in the (whole) target object instances"



Question: What is a train?

a) Rail

b) Train + Rail

c) Train ☺

Weakly Supervised Learning

Multiple vs single instance

"The same features are only present in the (whole) target object instances"

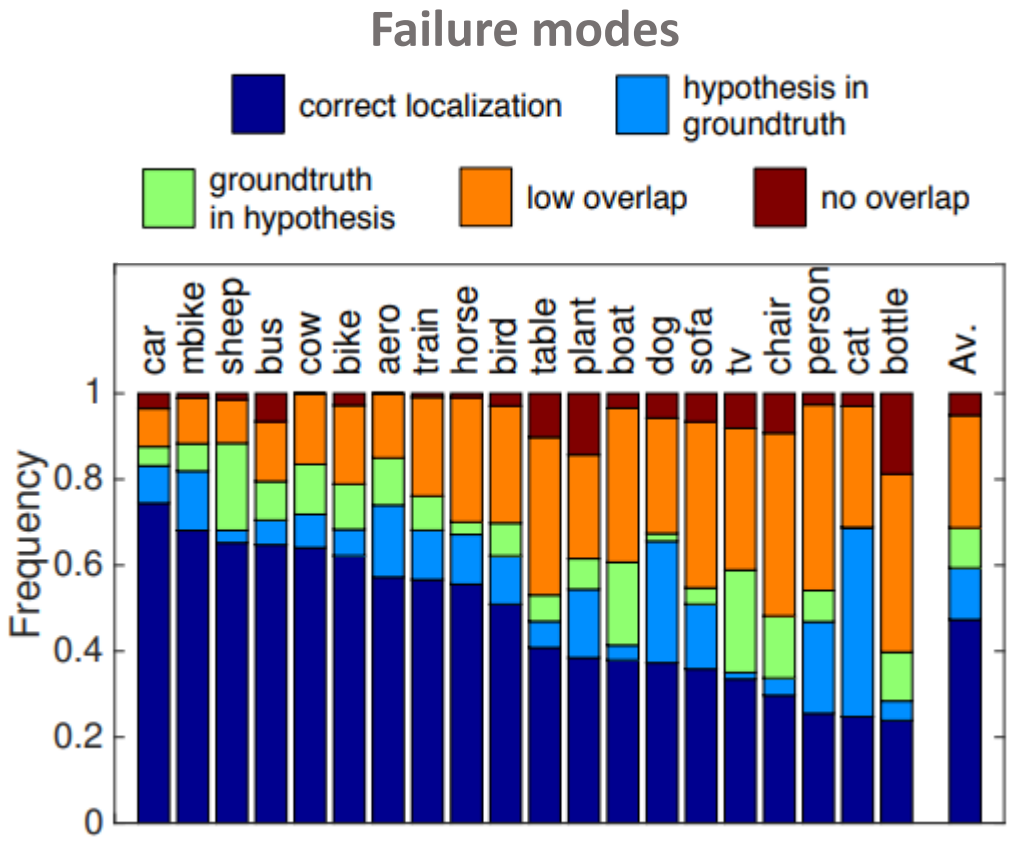


Question: What is a sheep?

a) 2 sheeps
b) 3 sheeps + grass
c) 1 sheep ☺

Weakly supervised object detection is
an ill-posed problem
due to

1. part/whole
2. foreground/background ambiguity



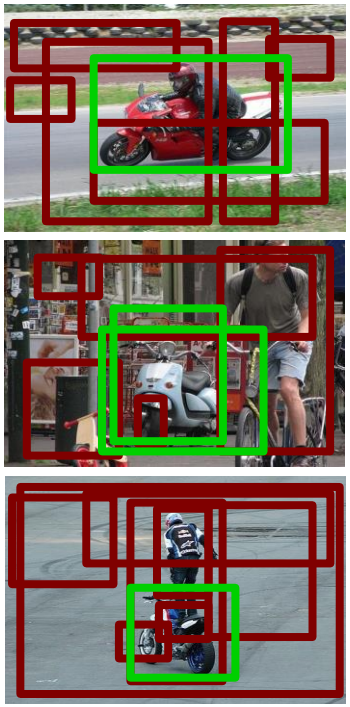
- **Inaccurate localization**: Most failures are in low overlap
- **Part vs whole**: Person, cat & dog face detection (hypothesis in gt)
- **Foreground vs background:** Sheep, boat and tv context detection (gt in hypothesis + low overlap + no overlap)

Image credit: Cinbis PAMI 17

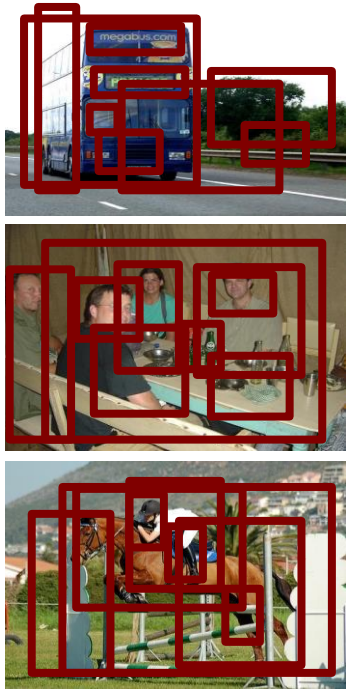
Principles for WSOD

Multiple-instance learning (MIL)

Positive bags



Negative bags



bags = images
instances = windows

Goals:

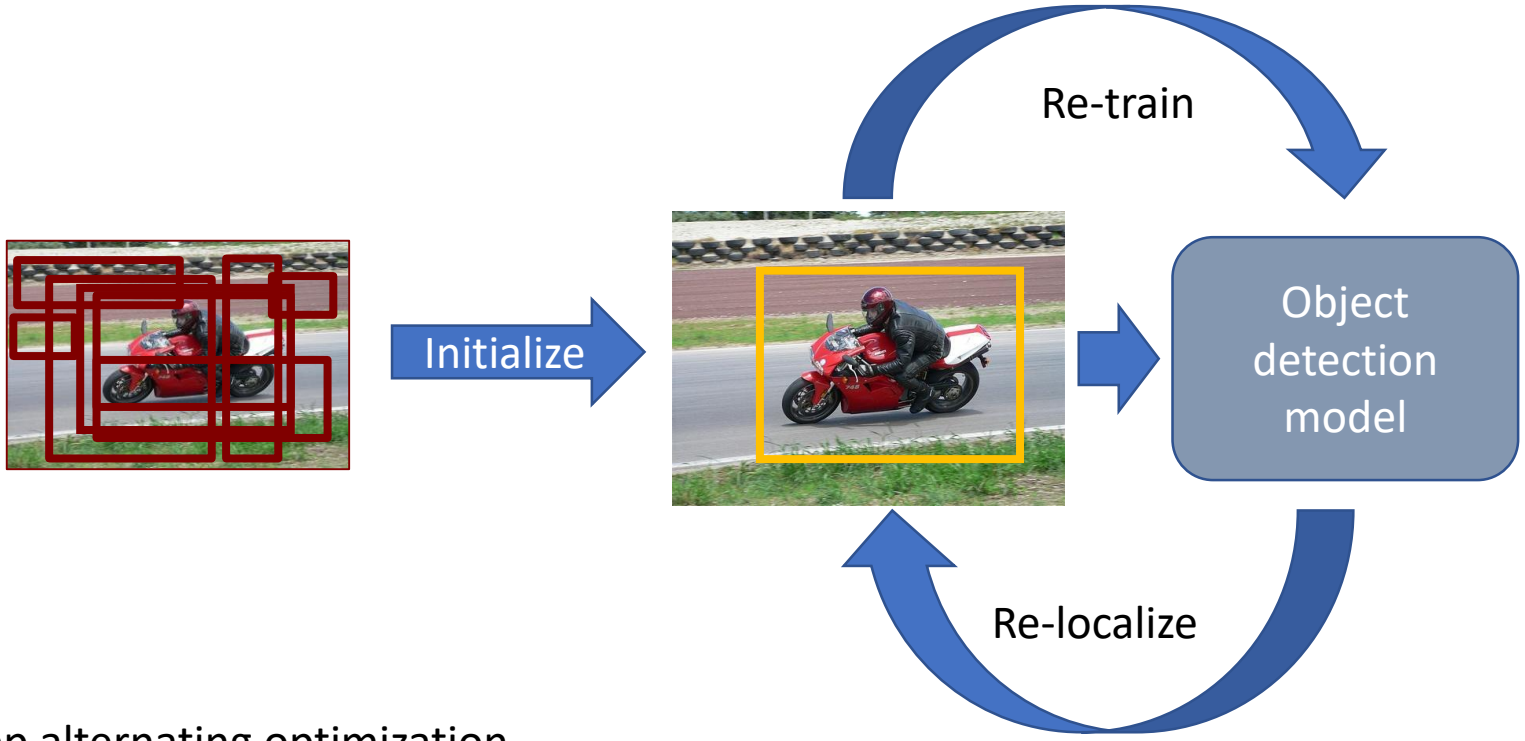
- find true positive instances
- train window classifier

MIL: [Dietterich 1997 AI]

MIL-WSOD: [Blaschko NIPS 10, Deselaers ECCV 10, Nguyen ICCV 09, Russakovsky ECCV 12, Siva ICCV 11, Siva ECCV 12]

Slide credit: Vittorio Ferrari

Standard MIL pipeline



Two step alternating optimization

- Re-train
- Re-localize

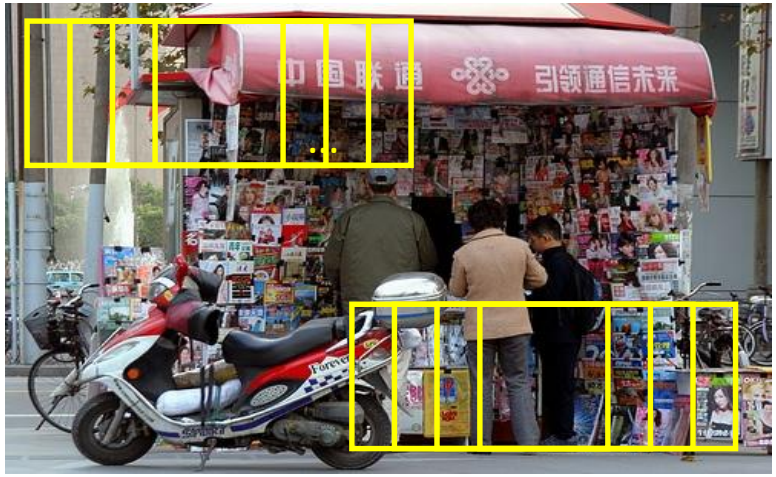
Challenges

- Sensitive to initialization (local minimum)
- Overfitting (locking) to predicted windows

How to generate bags?

Sliding windows

- >100k per image
- dense
- translations, scales and aspect-ratios (4D space)



[Chum CVPR 07, Nguyen ICCV 09, Pandey ICCV 11]

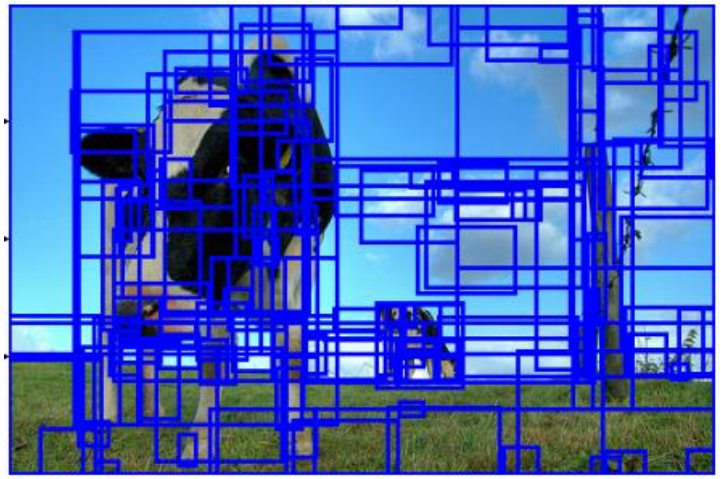
Object proposals

- Selective Search and Edge Boxes

[van de Sande ICCV 11, Dollar ECCV 14]

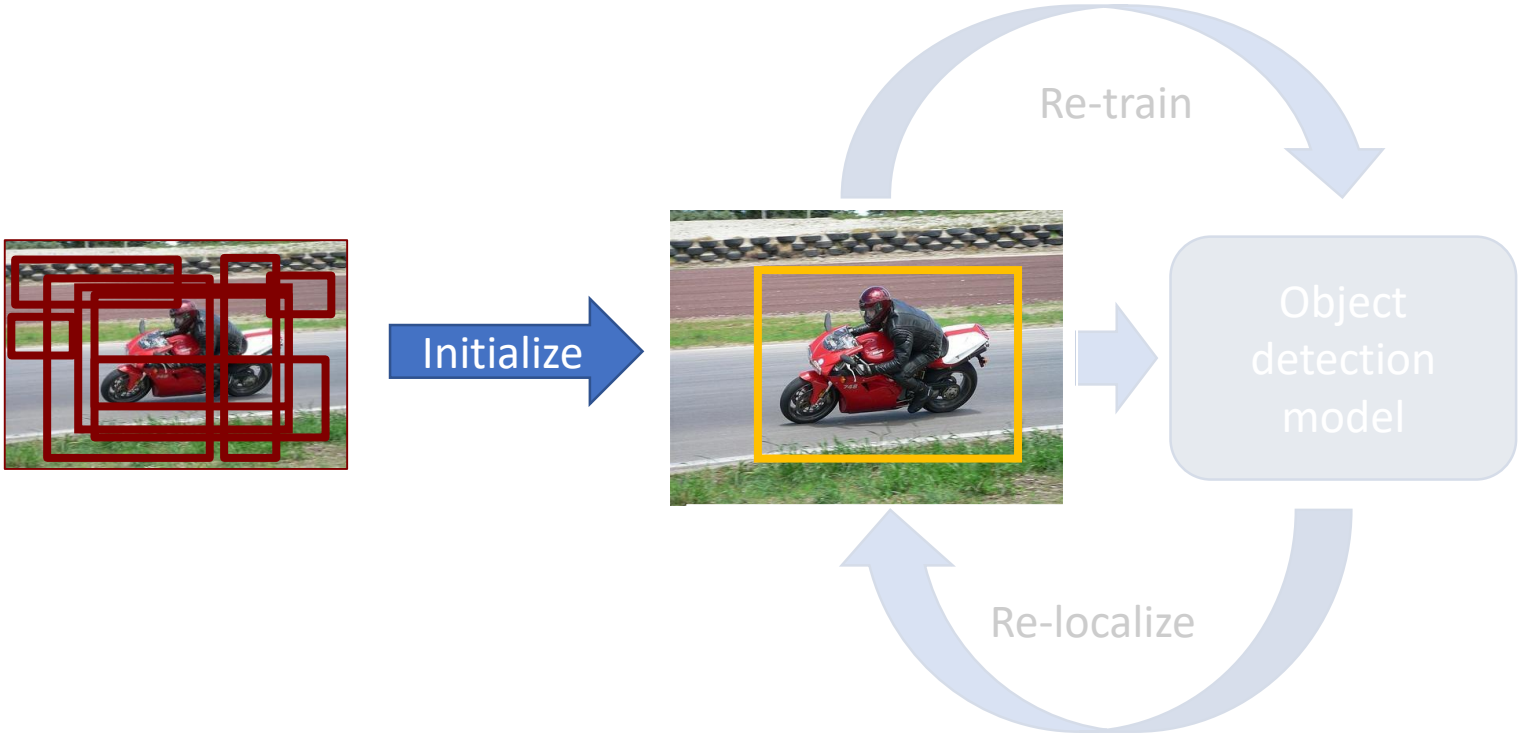
- ~2k per image
- Commonly used in WSOD

[Deselaers ECCV 10, Siva ICCV 11, Russakovsky ECCV 12, …]



Slide credit: Vittorio Ferrari

Standard MIL pipeline



Initialization

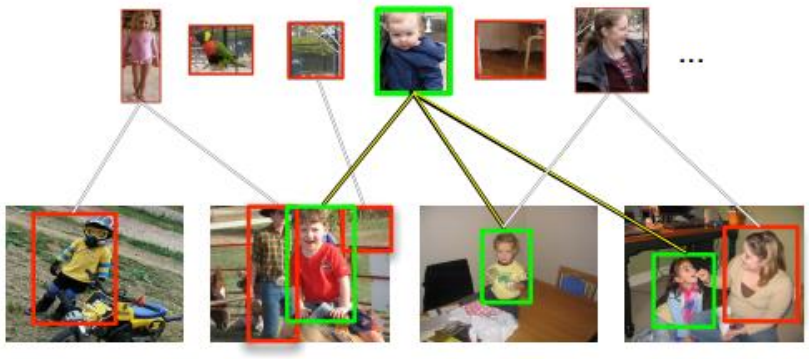
Simple strategies:

- Whole image

[Nguyen ICCV 09, Bilen BMVC 14]

- Whole image minus a margin

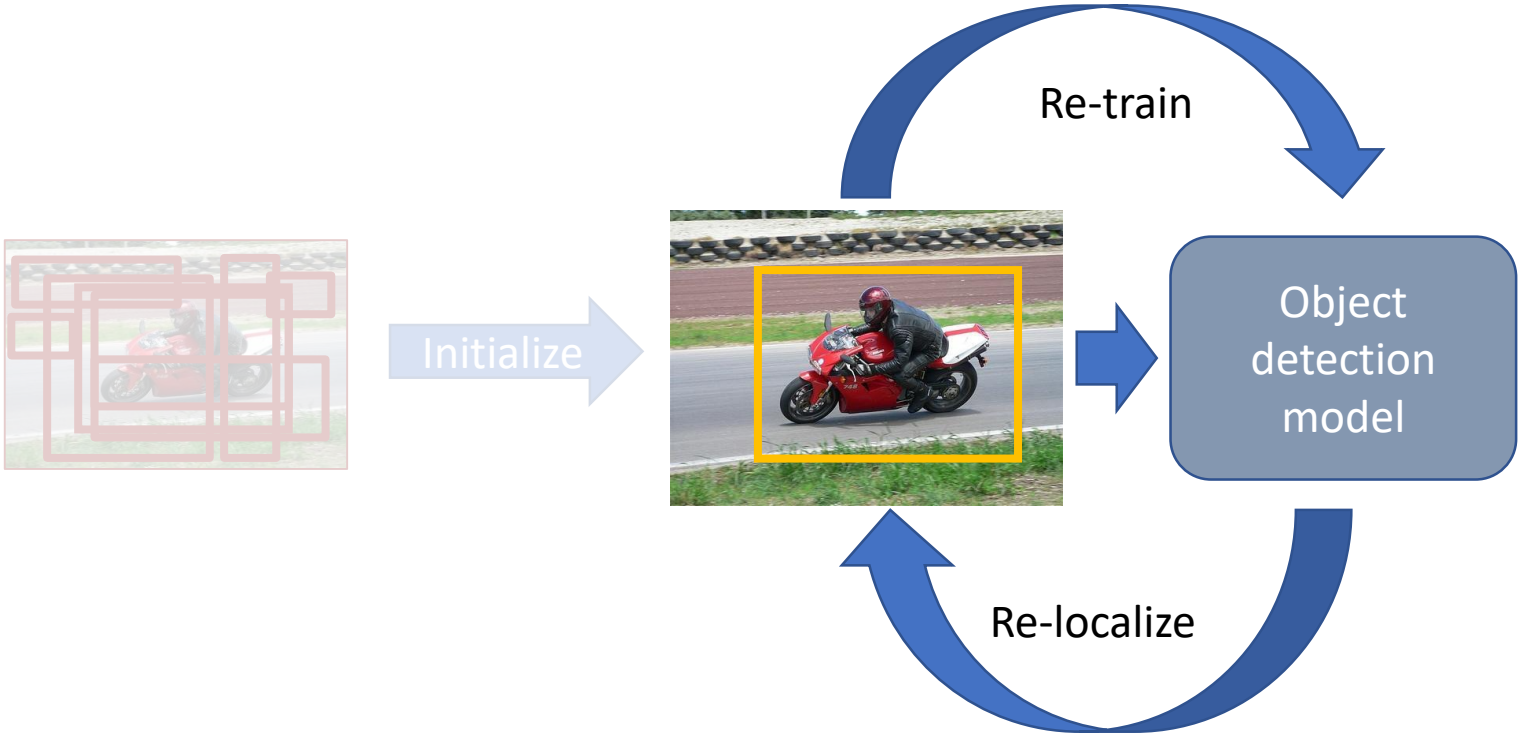
[Pandey ICCV11, Russakovsky ECCV12]



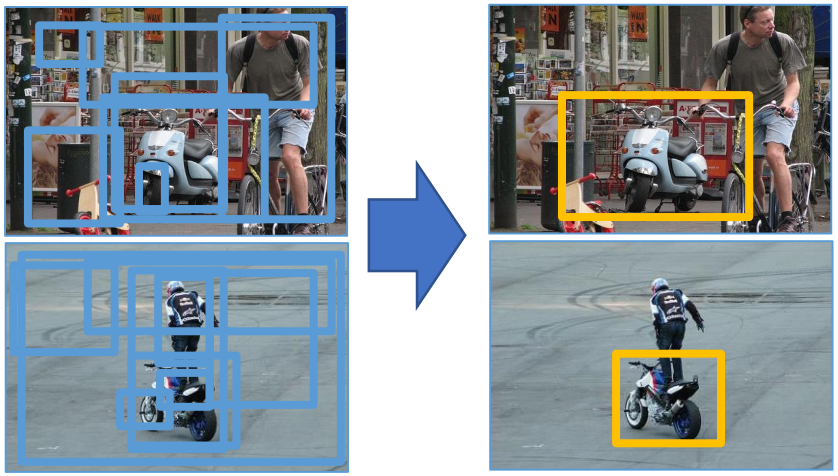
[Song ICML 14] constructs a graph to find initial boxes:

1. relevant (occur in many positive images)
2. discriminative (dissimilar to the boxes in the negative images)
3. complementary (capture multiple modes)

Standard MIL pipeline



Standard max formulation



$$\arg\max_{b} A(x_b)$$

Proposal — Appearance model — Features

- Pick the highest scoring window as positive
- Only one positive instance per image

[Nguyen ICCV 09, Bilen CVPR 14, Cinbis CVPR 14, Papadopoulos CVPR 16]

Hinge and cross-entropy loss

For positive images:

$$\max_b A(x_b) > \Delta \text{ (}\Delta\text{:margin) or } \log(\text{Softmax}(\max_b A(x_b)))$$

For negative images:

$$\max_b A(x_b) < -\Delta \quad \text{or} \quad \log(1-\text{Softmax}(\max_b A(x_b)))$$

Difference to supervised object detection (eg Faster-RCNN)

- Only 1 positive instance is used in each positive image
- No negative instances from positive image

[Nguyen ICCV 09, Bilen CVPR 14, Cinbis CVPR 14, Papadopoulos CVPR 16]

More robust optimization: Relaxing max operator

Hedge your bets on multiple proposals

Re-train

For positive images

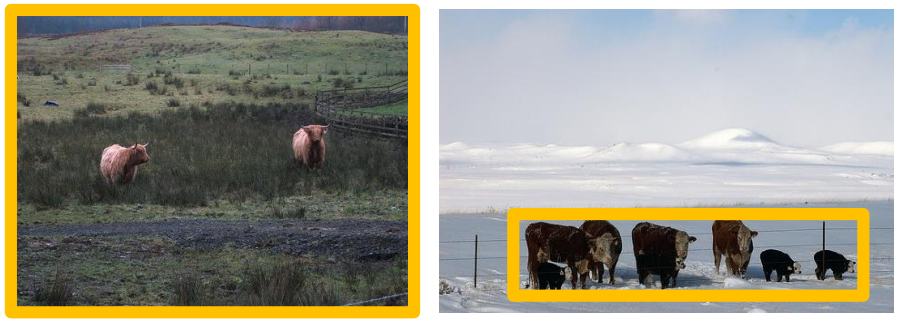
$$\log \sum_b \exp A(x_b^+) > \Delta$$

For negative images

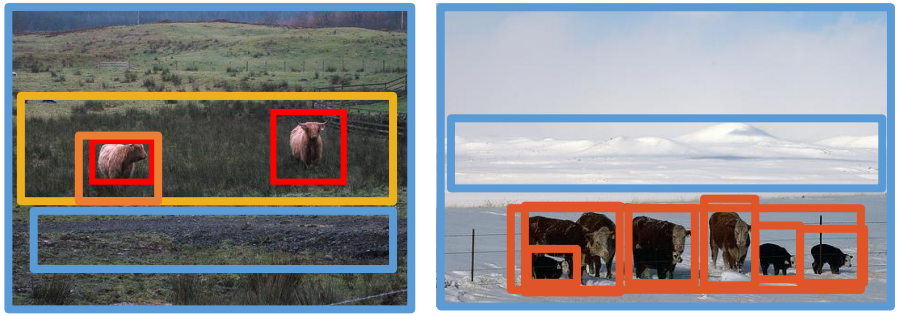
$$\log \sum_b \exp A(x_b^-) < -\Delta$$

- **Less sensitive to initialization**
- **Leverages multiple positive samples during training**

Max



Soft-max



[Bilen BMVC 14]

More robust optimization: Self-paced learning

- Inspired from Curriculum Learning
  [Bengio ICML 09]
- Start with easy samples, then consider hard ones in training

- Easiness measures for WSOD:
  - Selection of samples via confidence of max scoring window
    [Kumar NIPS 10]
  - Selection of window space by allowing smaller windows
    [Bilen IJCV 14, Shi ECCV 14]
  - Selection of samples via inter-category competition
    [Sangineto PAMI 17]

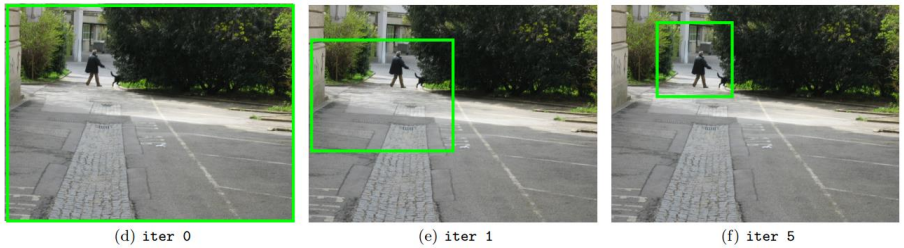


Image credit: [Bilen IJCV 14]

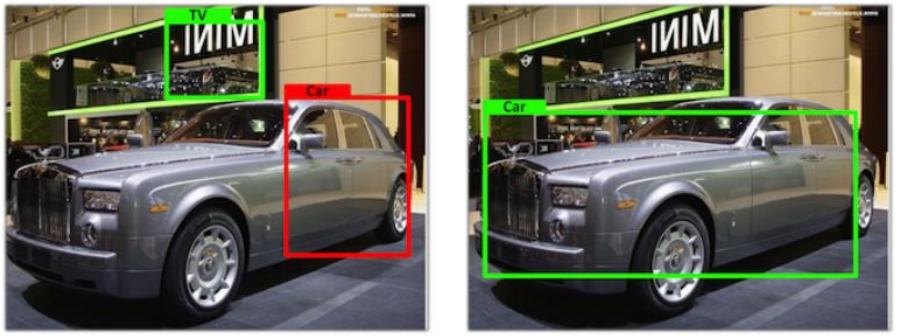
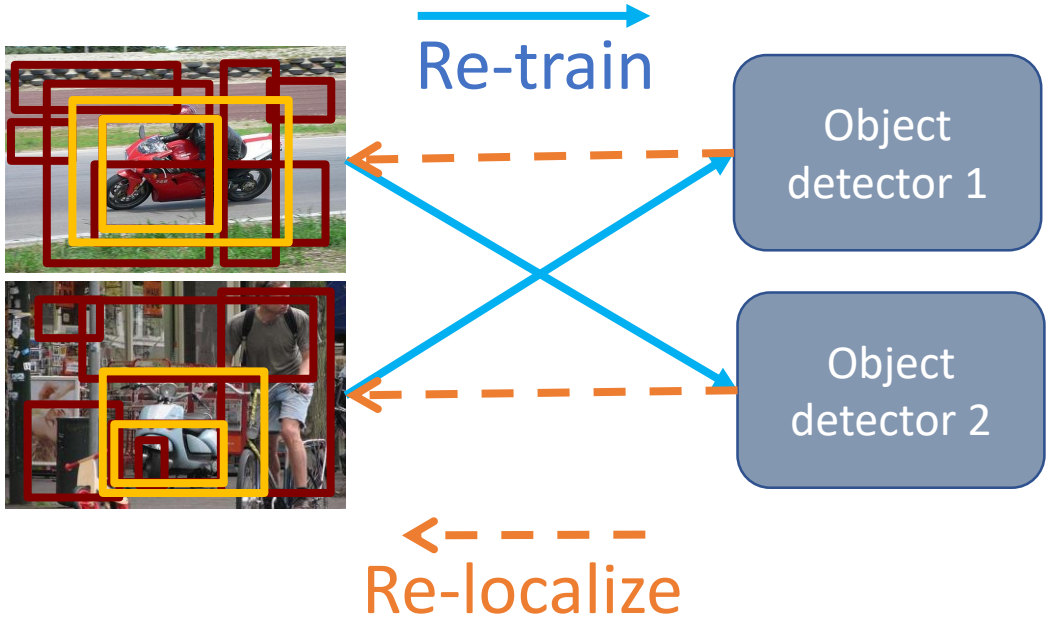


Image credit: [Sangineto PAMI 17]

More robust re-localization: Multifold MIL [Cinbis CVPR 14]

Problem: Detector overfits into the given proposal

Solution: Train using positive examples in all folds but k, and all negative examples



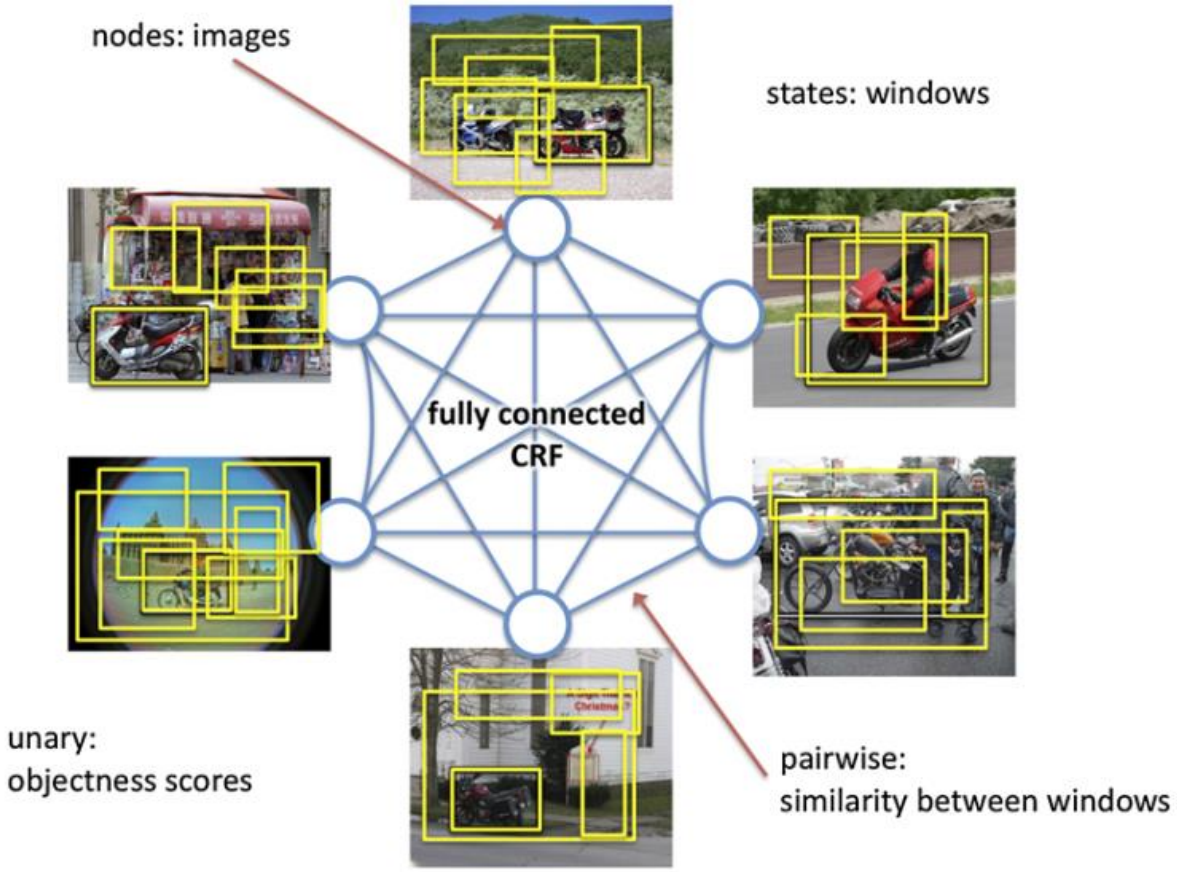
Pairwise similarity

Similarity between selected windows across positive images

☺ Less overfitting

☹ Expensive to optimize

☹ Ignores intra-class variation



[Chum CVPR 07, Deselaers ECCV 10, Siva ICCV 11, Bilen CVPR 15]

Image credits: [Deselaers ECCV 10]

Pairwise similarity

Sub-categories

Clustering via probabilistic latent Semantic Analysis (pLSA)

[Wang ECCV 14]

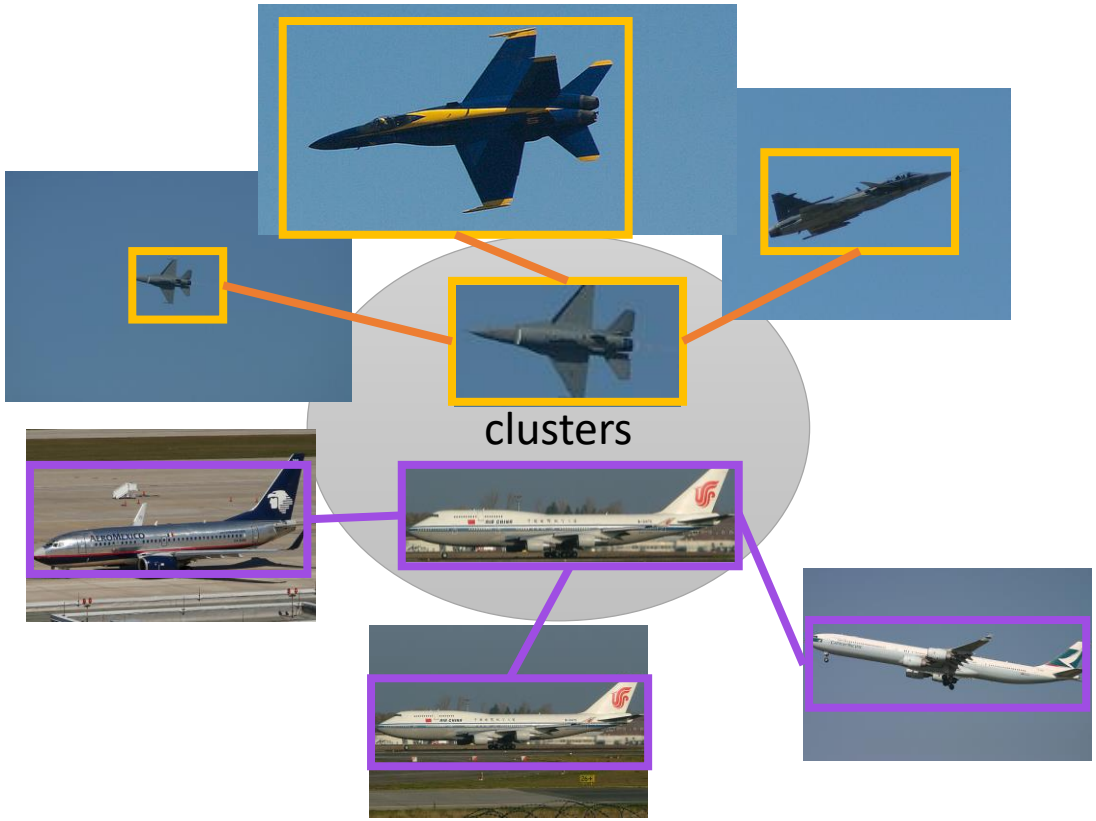
☺ Modeling intra-class variations

☹ Sensitive to number of clusters

Exemplars

Pick representative samples

☺ No need to set number of clusters



[Chum CVPR 07, Bilen CVPR 15]

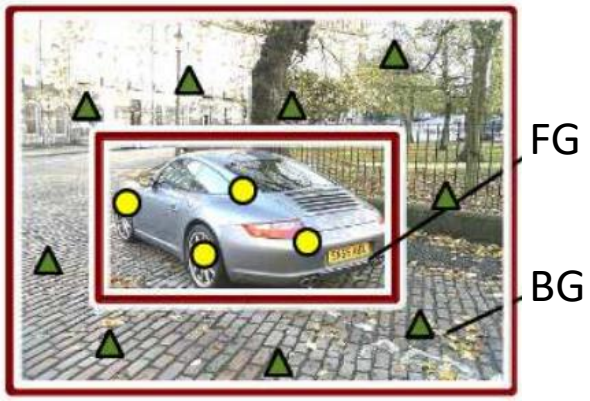
Image credit: [Bilen CVPR 15]

Additional cues: context

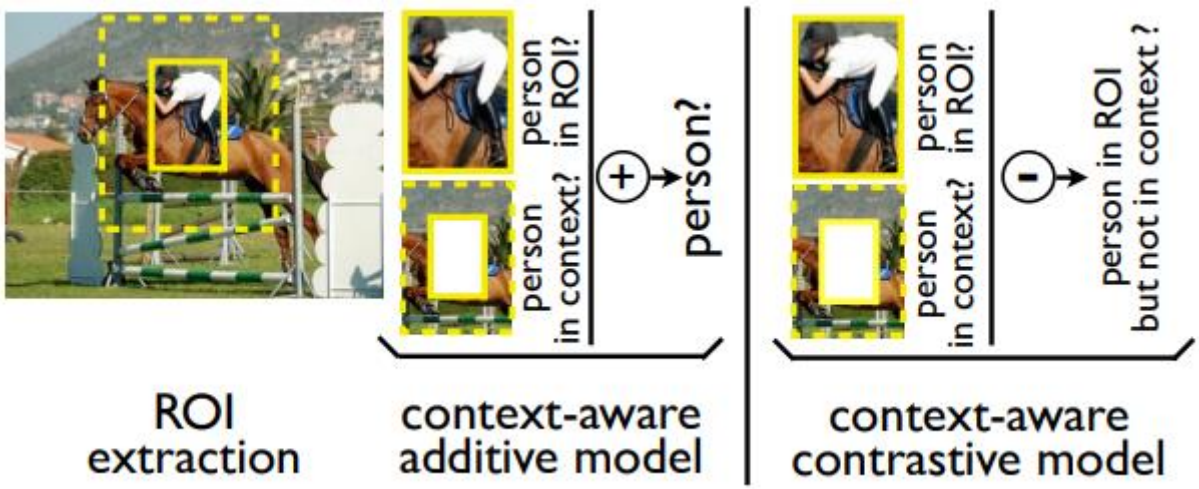
- Background provides contextual cues for recognition

[Russakovsky ECCV 12, Bilen CVPR 14, Kantorov ECCV 16]

- Better separation of foreground and background
- Additive: select a ROI that is semantically compatible with its context
- Contrastive: select a ROI that is outstanding from its context



[Russakovsky ECCV 12]



[Kantorov ECCV 16]

Additional cues: Objectness

Quantify how likely a window is to contain an object of *any* class

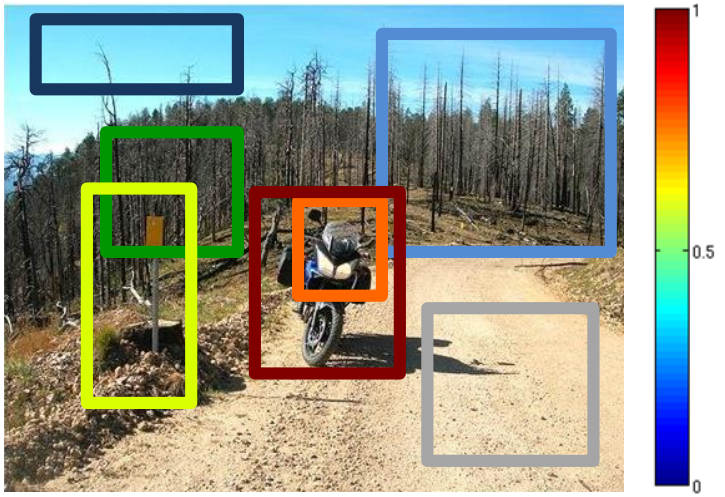
[Alexe CVPR 10, Uijlings IJCV 13, Zitnick&Dollar ECCV 14]

- Generates a small set of accurate object proposals
- Steers re-localization towards objects and away from background
- Pushes towards whole objects instead of sub-regions

$$argmax_b \lambda A(x_b) + (1-\lambda)\mathrm{Obj}(b)$$

Commonly used for weakly supervised object Localization

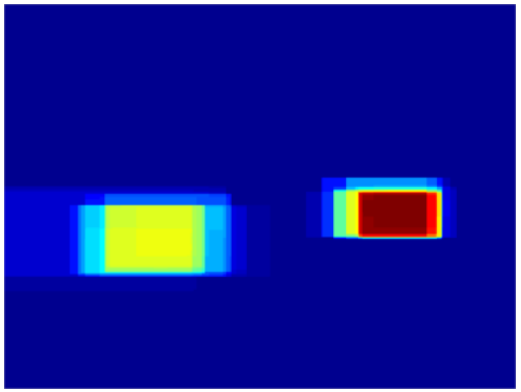
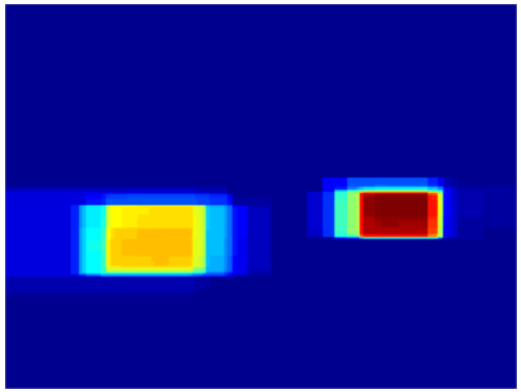
[Deselaers ECCV 10, Khan OAGMW 11, Siva ICCV 11, Guillaumin CVPR 12, Prest CVPR 12, Shapovalova ECCV 12, Shi BMVC 12, Tang CVPR 14, Wang ECCV 14, Jerripothula ECCV 16, Cinbis PAMI 16, Bilen CVPR 16]

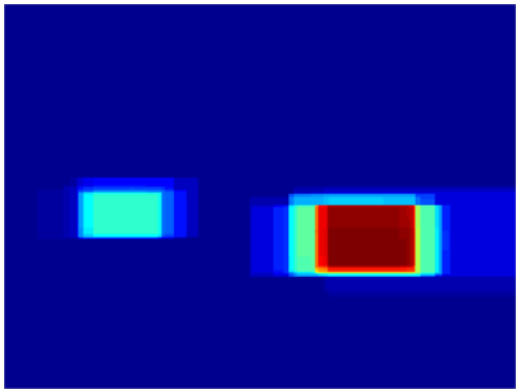
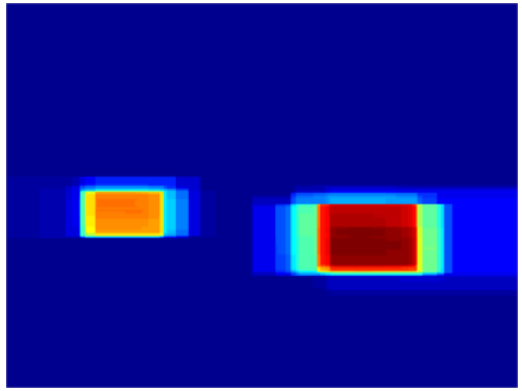


Slide credit: Vittorio Ferrari

Priors: Equivariance

The predicted output should be equivariant to image transformations

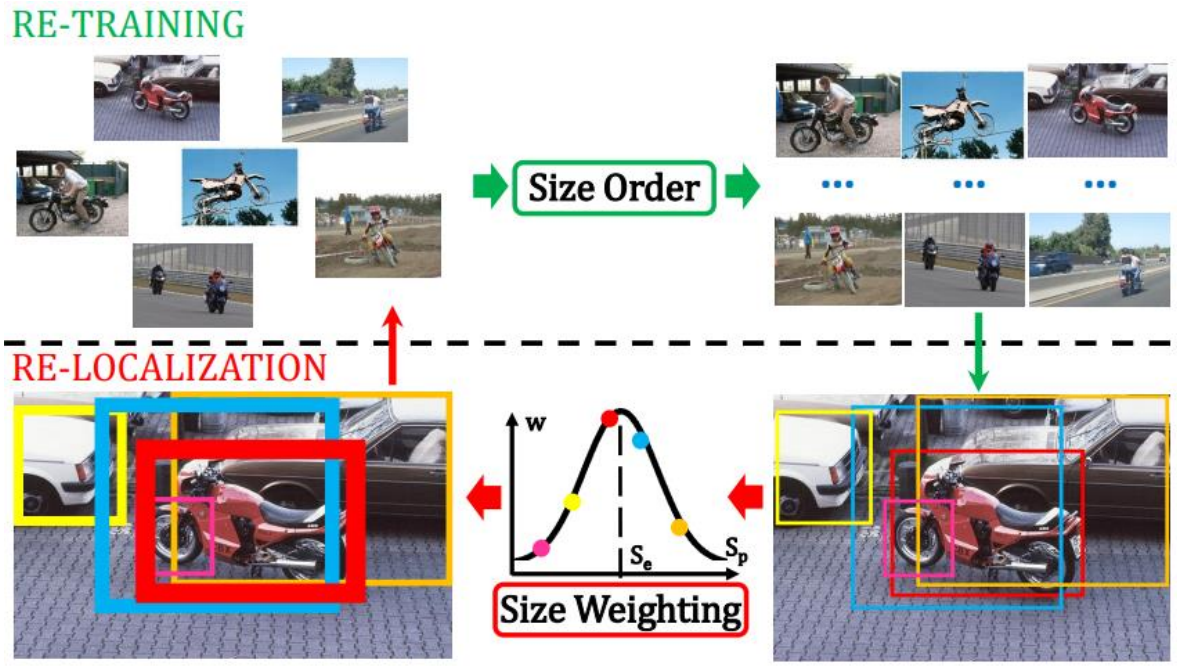



Before After

[Bilen BMVC 14]

Priors: Scale [Shi ECCV 16]

- Curriculum learning (bigger objects down to smaller ones)
- Weight object proposals according to estimated size
- Requires training a size estimator from a small set



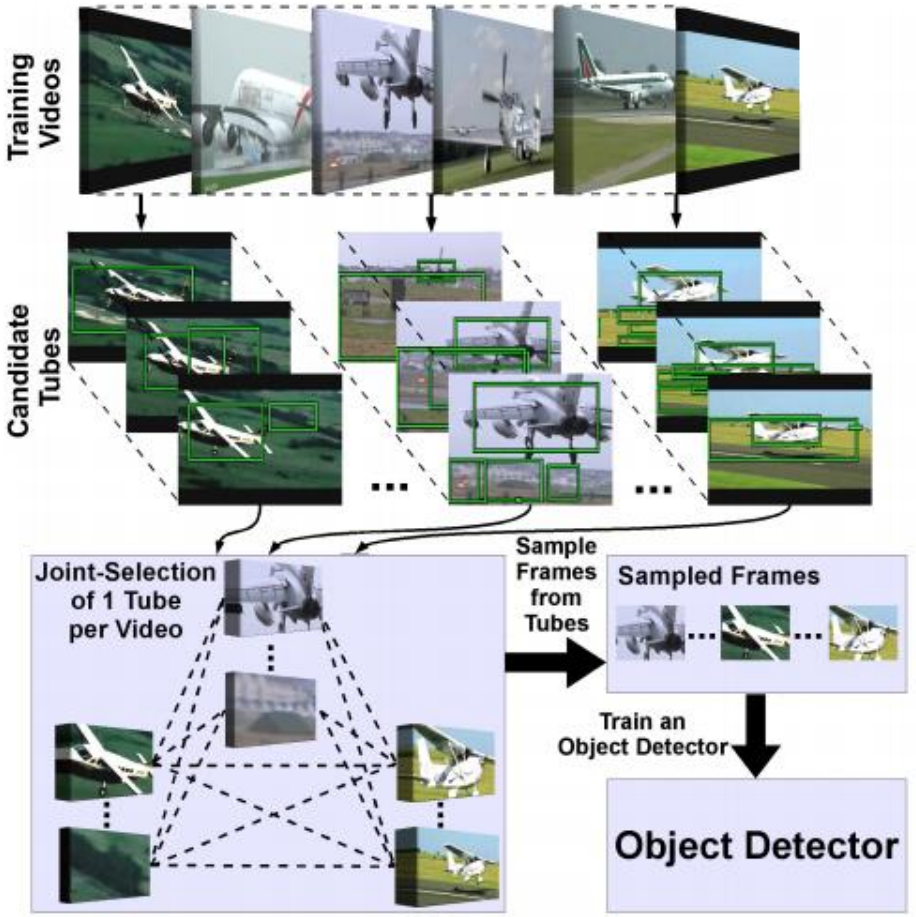
[Shi ECCV 16]

Priors: Motion

1. extracts candidate spatio-temporal tubes based on motion segmentation
2. selects one tube per video jointly over all videos
3. uses objectness to reduce object candidates
4. applies a domain adaptation method to address the video/image gap

☺ Motion cues for object boundaries

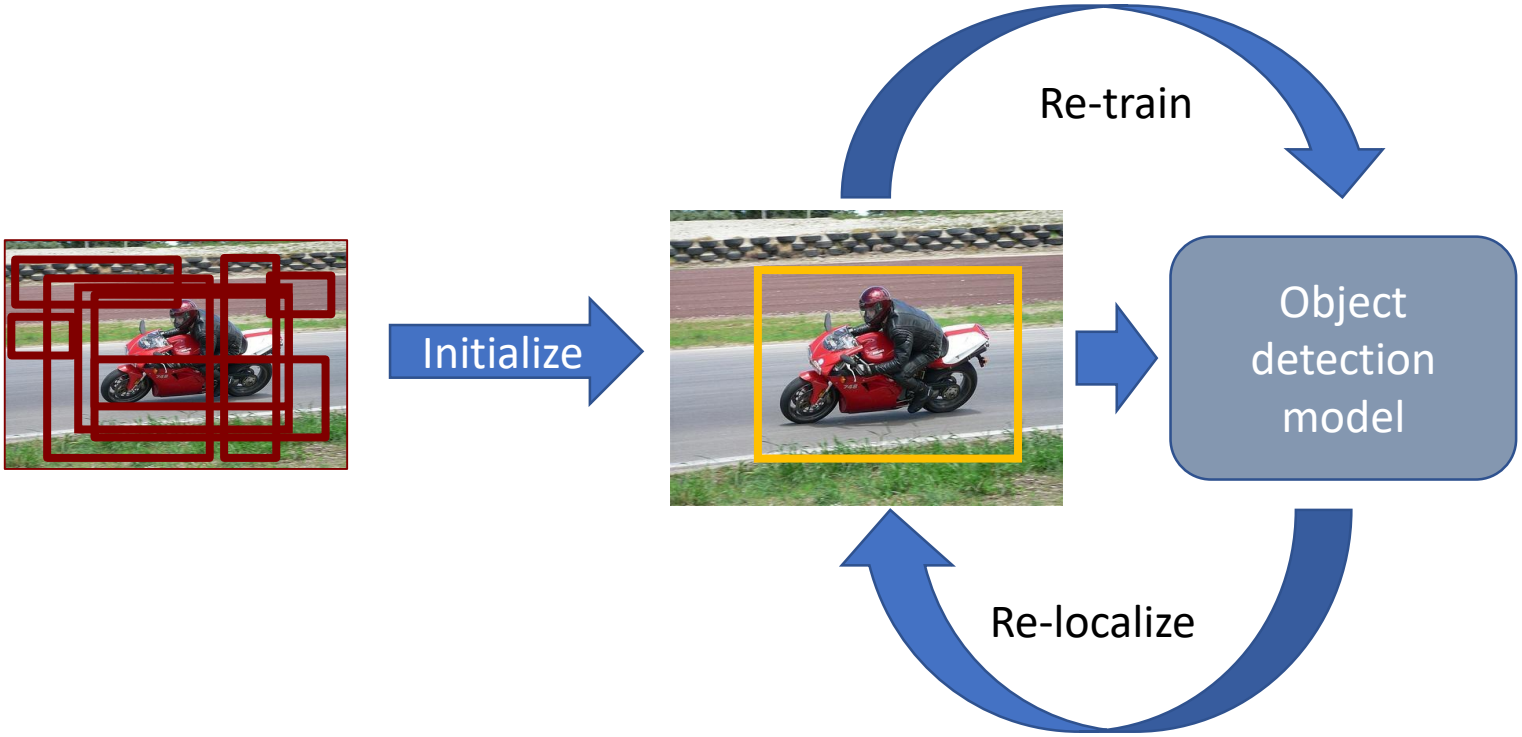
☹ Noisy data



[Prest CVPR 12, Tang CVPR 13, Joulin ECCV 14, Kuznetsova CVPR 15, Liang ICCV 15, Liang ICCV 15, Kalogeiton PAMI 15]

Image credit: [Prest CVPR 12]

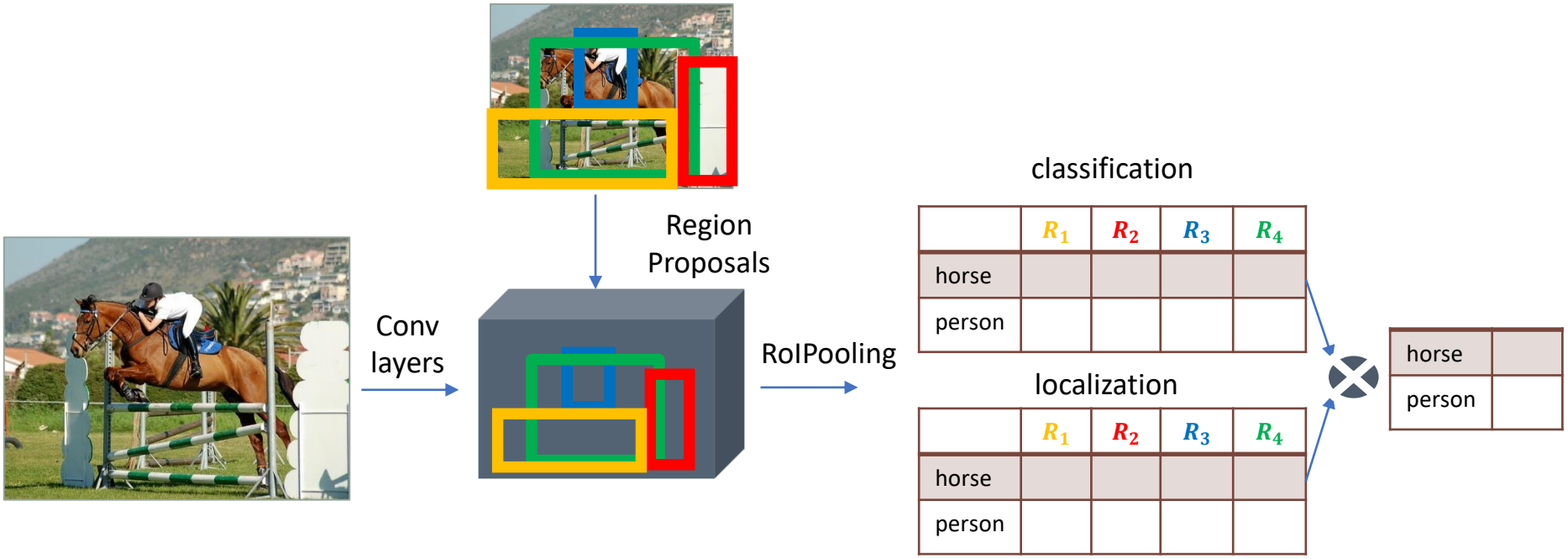
Standard MIL pipeline



Deep and End-to-End WSOD

Weakly Supervised Deep Detection Network (WSDDN) [Bilen CVPR 16]

- Previous work uses CNN as black-box feature extractor
- First end-to-end WSOD method, finetunes a pre-trained deep network
- Uses RoI-Pooling [Girshick ICCV 15]
- Two stream architecture
  - Classification: Associates each region with a class score
  - Localization: Compares regions for each class



WSDDN

☺ State-of-the-art results with VGG-F (~AlexNet) (62% of supervised)

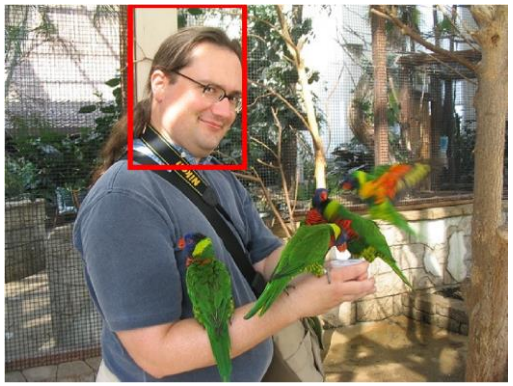
☺ End-to-end learning + No custom deep learning layers

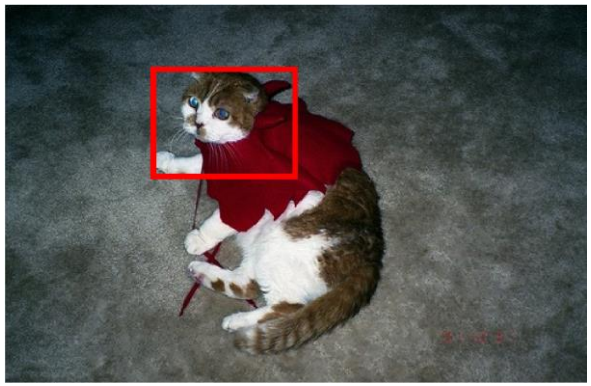
☹ Detects discriminative parts rather than whole object extent

☹ Groups multiple instances together

Limitations (compared to the supervised case, eg Faster RCNN):

- It does not leverage multiple positive instances in positive images
- It does not exploit negative instances in positive images
- It does not learn bounding box regression





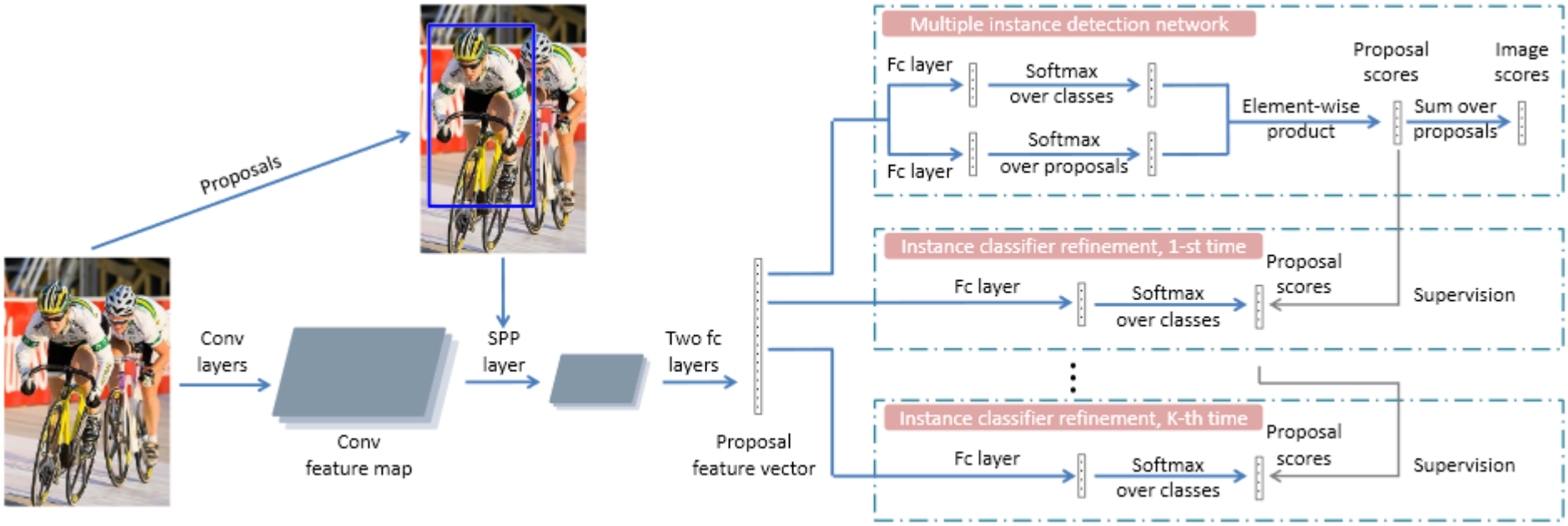
[Bilen CVPR 16]

Online Instance Classifier Refinement (OICR) [Tang CVPR 17]

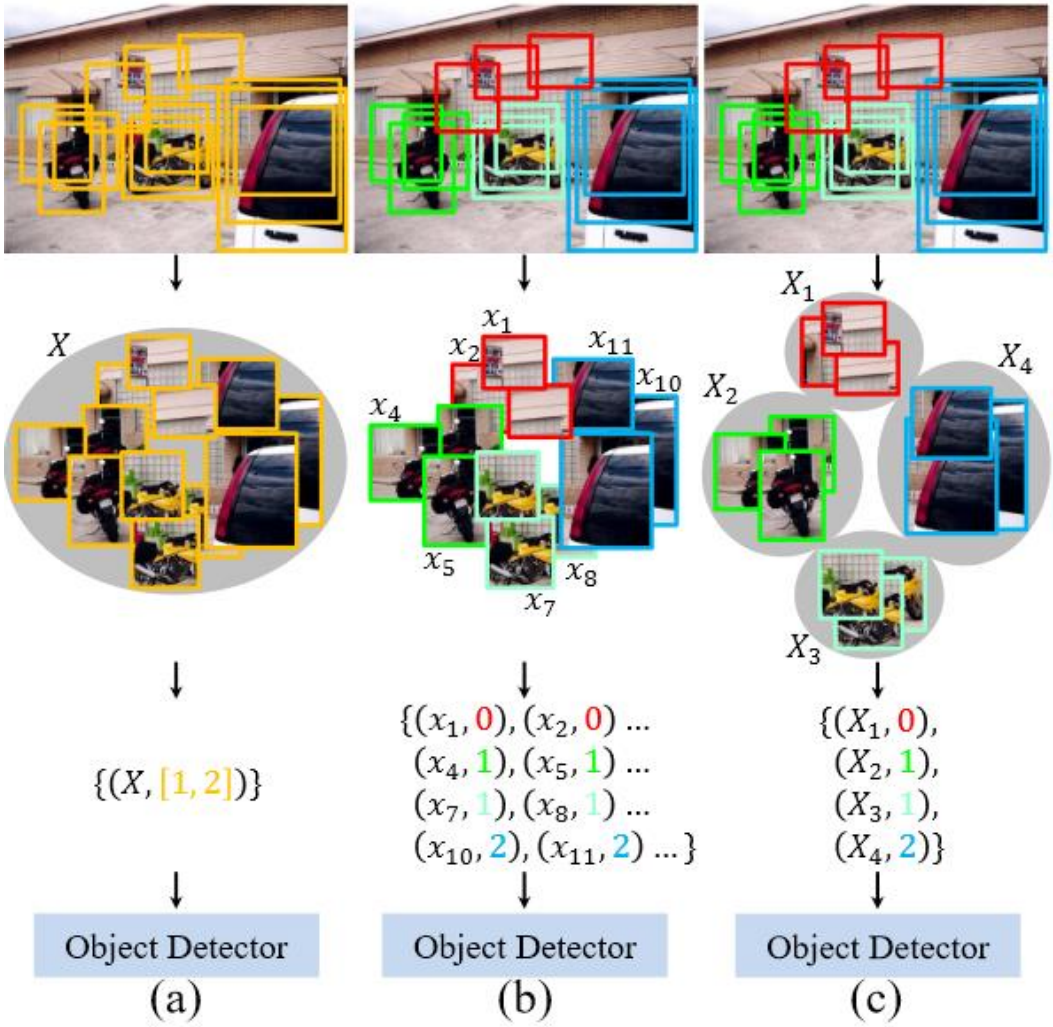
WSDDN learns to classify images based on implicit instance scoring

OICR learns to classify instances

- Initially uses WSDDN to set instance labels
- Iteratively refines them by using an instance classifier



Proposal Cluster Learning (PCL) for WSOD [Tang PAMI 18]



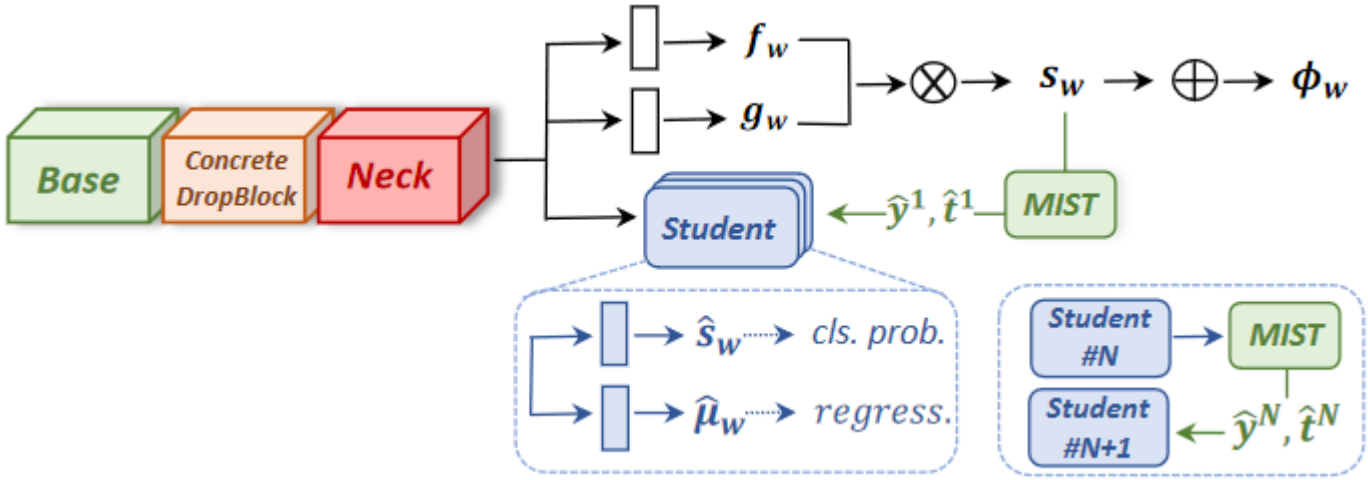
Observation: There can be multiple positive instances in an image

OICR considers only the highest scoring proposal + surrounding ones as positives

PCL

- builds an undirected graph of top- ranking proposals
- iteratively selects vertices which have most connections to be the cluster centres

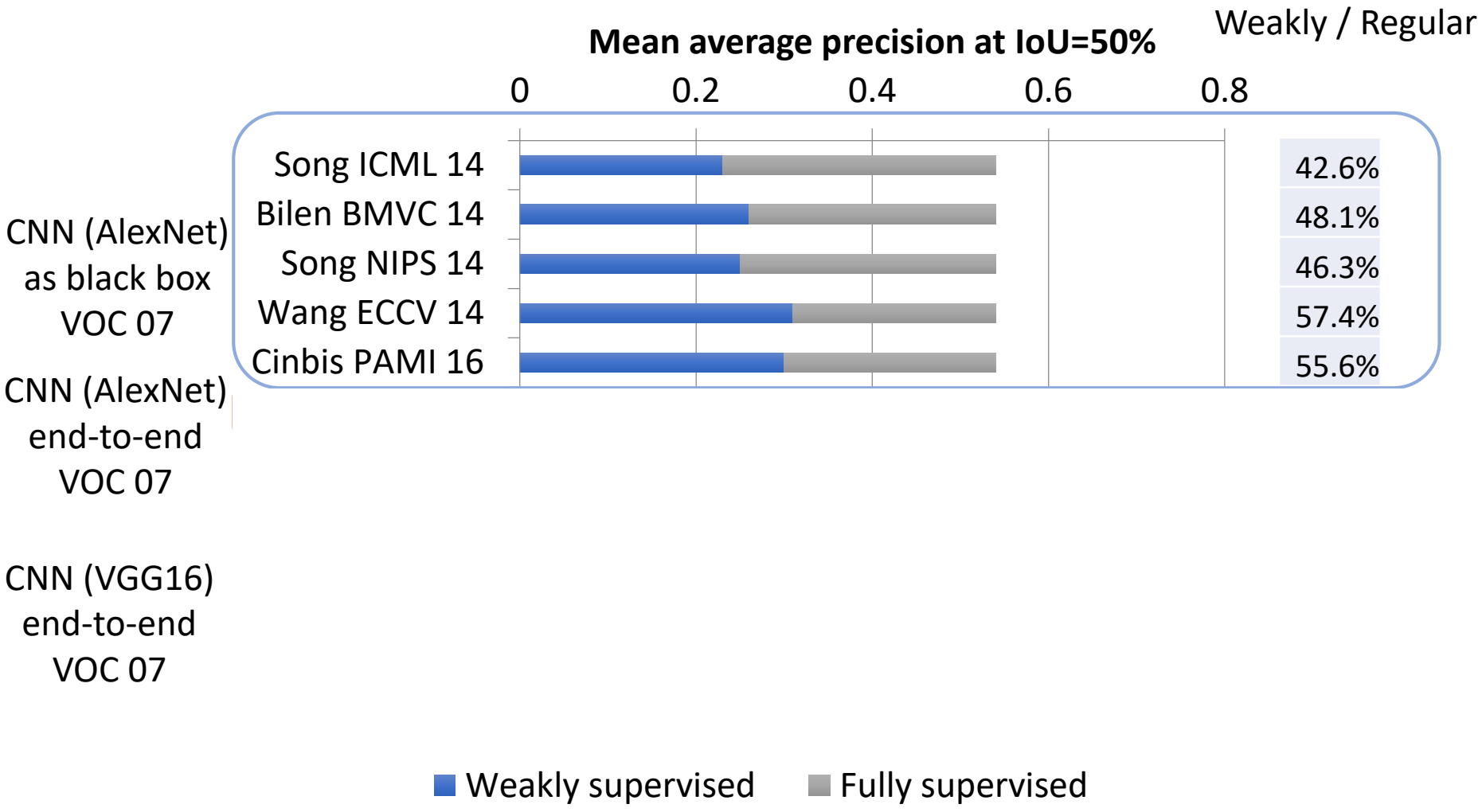
Instance-Aware, Context-Focused, and Memory-Efficient WSOD [Ren CVPR 20]



- Builds on **WSDDN** and **OICR**
- Picks top p % of the highest scoring proposals and surrounding ones as positives
- Introduces ConcreteDropBlock: a data-driven and parametric drop-out to drop the most relevant regions
- Adversarial optimization

$$w^*, \theta^* = \arg\min_{w} \max_{\theta} \sum_{I} \mathcal{L}_{\text{img}}(w, \theta) + \mathcal{L}_{\text{roi}}(w, \theta)$$

VOC 07 Performance at test time



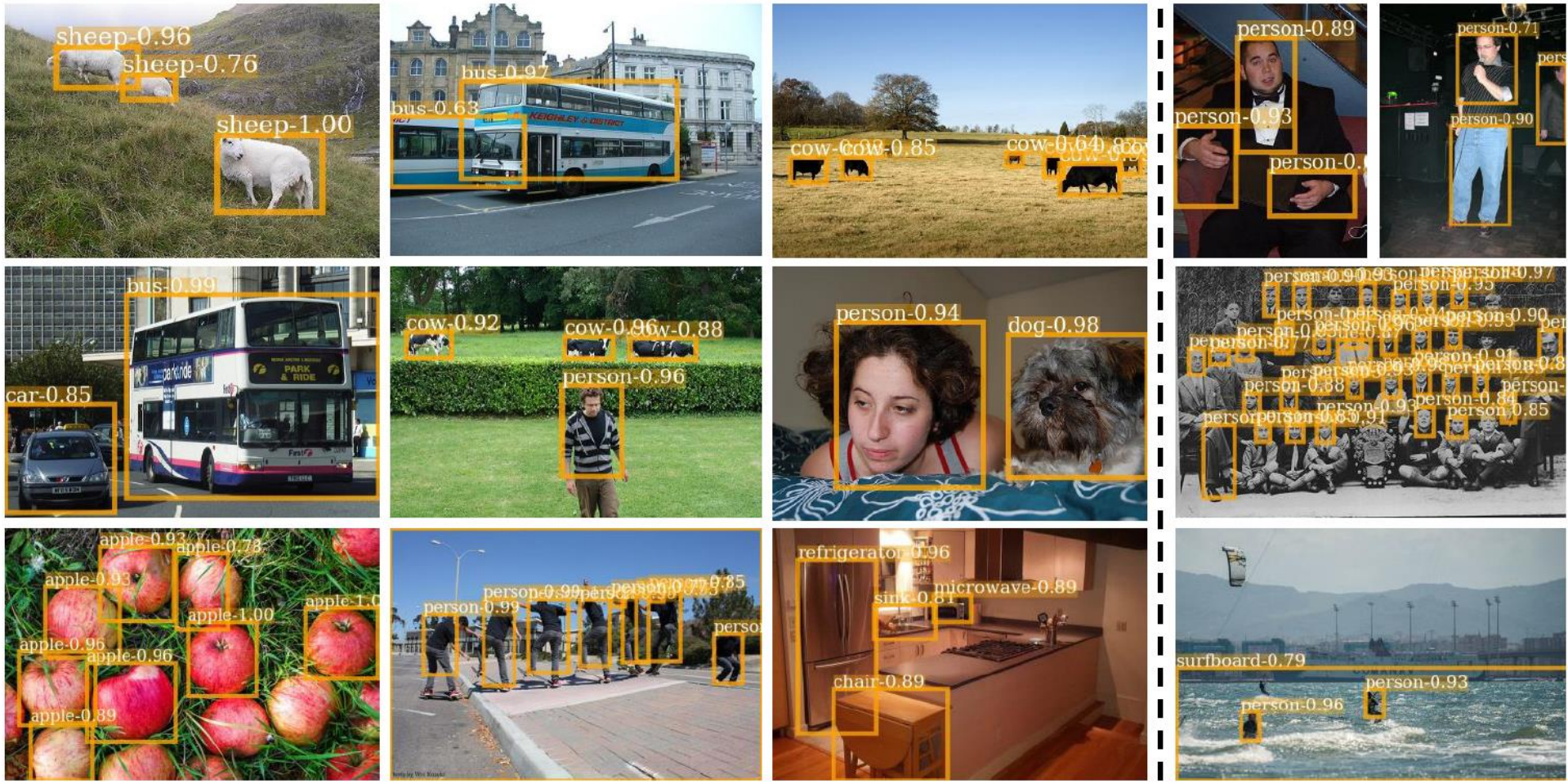
Daring progression!

Visualizing state-of-the-art WSOD results

[Ren CVPR 20]

Successful cases

Failure cases



Summary on weakly supervised object detection

- WSOD is an **ill-posed problem** and ambiguous due to part/whole and co-occurence issues.
- WSOD training is **sensitive to initialization and prone to overfitting**.

- Remarkable progress due to
  - objectness and region proposals,
  - incorporating prior knowledge and additional cues,
  - end-to-end learning.

- Addressing the inherent ambiguity requires better **objectness measures** and stronger **additional cues**.